Secure Wide Area Network Access to CMS Analysis Data Using the Lustre Filesystem

D Bourilkov[1], P Avery[1], M Cheng[1], Y Fu[1], B Kim[1], J Palencia[2], R Budden[2], K Benninger[2], J L Rodriquez[3], J Dilascio[3], D Dykstra[4] and N Seenu[4]

[1] University of Florida, Gainesville, Fl, USA
[2] Pittsburgh Supercomputing Center, Pittsburgh, PA, USA
[3] Florida International University, Miami, Fl, USA
[4] Fermi National Accelerator Laboratory, Batavia, Il, USA

E-mail: `bourilkov,avery,cheng,yfu,bockjoo@phys.ufl.edu`
`josephin,rbudden,benninge@psc.edu`
`jrodrig,jdilascio@fiu.edu`
`dwd,seenu@fnal.gov`

Abstract. This paper reports the design and implementation of a secure, wide area network (WAN), distributed filesystem by the ExTENCI project (Extending Science Through Enhanced National CyberInfrastructure), based on the Lustre filesystem. The system is used for remote access to analysis data from the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC), and from the Lattice Quantum ChromoDynamics (LQCD) project. Security is provided by Kerberos authentication and authorization with additional fine grained control based on Lustre ACLs (Access Control List) and quotas. We investigate the impact of using various Kerberos security flavors on the I/O rates of CMS applications on client nodes reading and writing data to the Lustre filesystem, and on LQCD benchmarks. The clients can be real or virtual nodes. We are investigating additional options for user authentication based on user certificates.

1. Introduction

The Lustre [1] parallel, distributed filesystem is used in some of the world's largest and most complex computing environments. It provides high performance, scaling to tens of thousands of nodes and petabytes of storage with groundbreaking I/O and metadata throughput. Lustre 2 [2] provides support for OEL 5, RHEL 5, SLES 10 and 11 (client only), and Fedora 11 (client only). This release series offers a number of significant features and enhancements, including Changelogs, Commit on Share, Lustre rsync, and Size-on-MDS.

Kerberos [3] is a network authentication protocol designed to provide a strong authentication mechanism for client-server applications through secret-key cryptography. Additionally, when users request access to the GSS-protected filesystem, Kerberos nominally performs user and service credential checks over the network. Within Kerberos, realms are created, and principals (systems and users) are placed in secure realms where the principals authenticate to other principals within the same realm.

Principals in one realm can also authenticate to principals in other realms to achieve cross-realm security. The capability for cross-realm authentication is a necessity in the insecure wide area network. Other security standards such as OpenID and InCommon are more oriented towards web applications. Globus, which uses X509 certificates, has similar capability for cross-realm, federated user ID but does not provide inherent Lustre filesystem security. To interoperate with the public key certificates and enable users to authenticate in a Kerberos realm using their X509 certificates, PKINIT is enabled within Kerberos.

The incorporation of Kerberos into Lustre extends the authentication to all the Lustre components and user access to the filesystem. The Lustre network can operate with RPC message and bulk data protection enforcing data checksum, privacy, or integrity. Furthermore, secure, distributed Object Storage Targets (OST) and OST pools decentralize the Lustre storage across organizations for faster local I/O.

We build a test bed for using Lustre over the WAN for two high energy applications. The CMS experiment [4] at CERN is a general purpose detector at the LHC, located in the vicinity of Geneva, Switzerland, which studies the collisions of protons and nuclei in the multi Tera electron Volt (TeV) range. Lattice Quantum ChromoDynamics is a computer intensive study of the confinement of quark-gluon plasma via the lattice gauge theory formulated on a grid, or lattice of points in spacetime. The discretized formulation of QCD [5] rather than continuous spacetime allows to explore the highly nonlinear nature of the strong nuclear force.

2. Lustre Test Bed

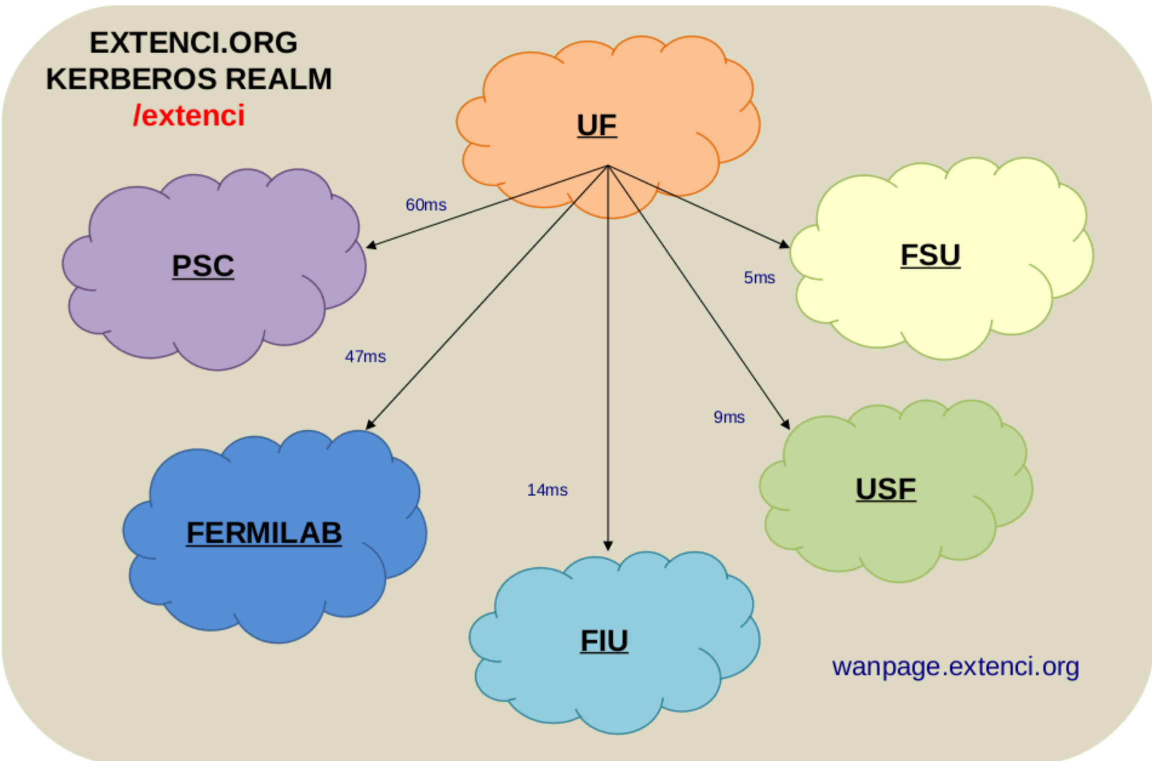
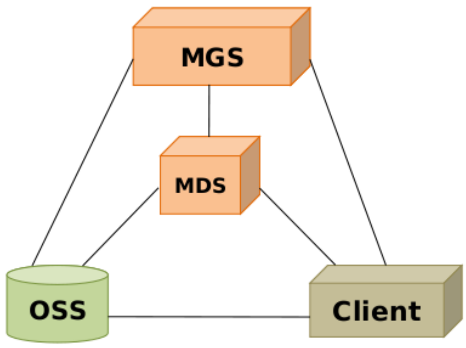


Figure 1. ExTENCI Lustre WAN General Overview. The Kerberos realm EXTENCI.ORG was established to create the secure Lustre network that only authorized systems and users can access. UF manages the KDC, metadata and OSS storage servers.

Authenticated Lustre Components



FLAVOR	AUTH	RPC MESSAGE PROTECTION	BULK DATA PROTECTION
lctl conf_param extenci.srpc.flavor.default = **krb5n**			
null		NULL	NULL
KRB5n	GSS/krb5	NULL	checksum(adler32)
KRB5a	GSS/krb5	PARTLY INTEGRITY	checksum(adler32)
KRB5i	GSS/krb5	INTEGRITY	integrity(sha1)
KRB5p	GSS/krb5	PRIVACY	privacy(sha1/aes128)

lctl conf_param extenci.srpc.flavor.**tcp0**=krb5n
extenci.srpc.flavor.**tcp1**=null
extenci.srpc.flavor.default.**cli2ost**=krb5i
extenci.srpc.flavor.default.**mdt2mdt**=null
extenci.srpc.flavor.default.**mdt2ost**=krb5i
mgs.srpc.flavor.default=krb5p

- Ease in bringing up secure lustre components
- Kerberos infrastructure is **NOT** required
- Each system is given a **UNIQUE** keytab
- Seconded by firewall (becomes **optional**)

Figure 2. ExTENCI Lustre WAN Detailed Overview. The different Kerberos flavors studied in this paper are described.

The basis components of a Lustre filesystem are:

- MDS: meta-data server
- OSS: object storage server
- OST: several object storage targets, organized in pools.

Files can be stored on a single OST, or striped over several OSTs. Kerberos is incorporated into the source development Lustre code by enabling GSS-support in the builds.

Figures 1 and 2 show a simple and a more detailed overview of the ExTENCI Lustre WAN filesystem, respectively. The Key Distribution Center (KDC) and the Lustre servers are based at the University of Florida (UF), an Open Science Grid (OSG) Tier2 site. The secure filesystem is accessed by other Tier3 universities such as Florida International University (FIU) and Florida State University (FSU), by Fermilab, a Tier1 site, and by PSC [6]. Virtual Lustre clients running XEN, VMware, VirtualBox and KVM mount the `/extenci` filesystem after being authorized and granted unique keytabs by UF. Figure 3 shows the detailed layout of the distributed OST filesystem, with pools located at UF with a fast 10 Gbps connection to the OSS server, and at Fermilab over the WAN. The total size of the storage pools is 60 TB.

3. Performance Tests and Results

The impact of Kerberos flavors on I/O rates over the local or wide area networks is shown in Figures 4 and 5. On the LAN, where the latencies are in the sub-millisecond range, the impact is visible, but not limiting. The effects are much more pronounced when accessing files over the

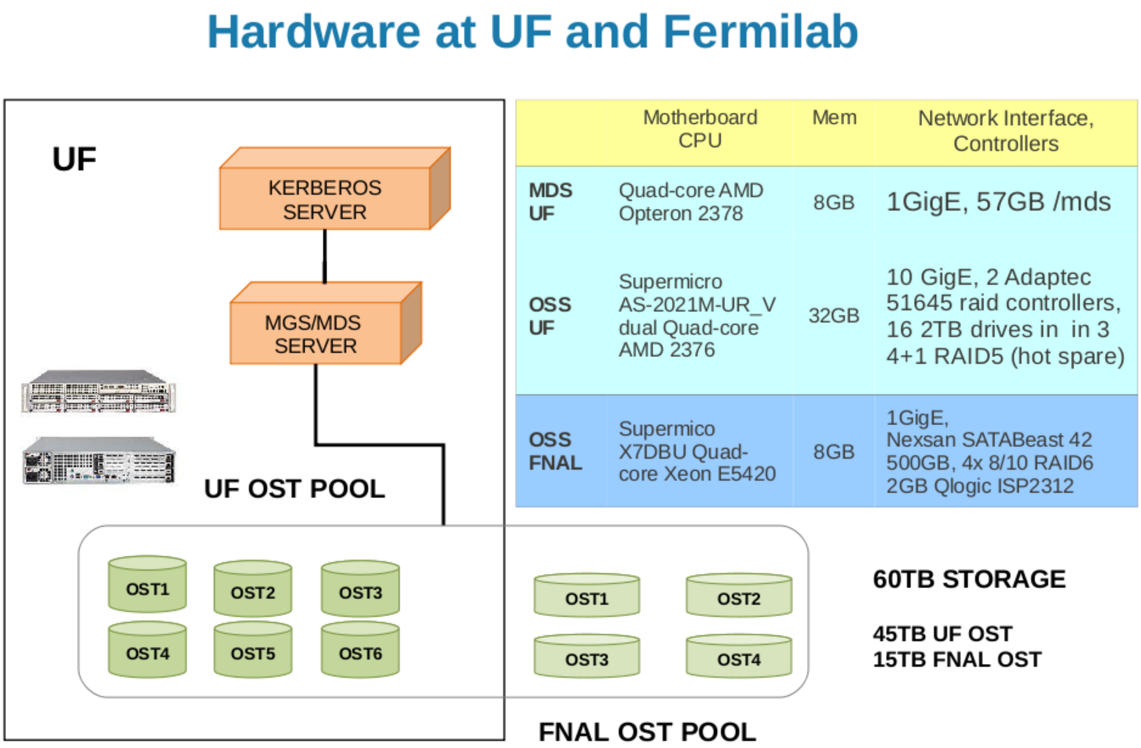


Figure 3. ExTENCI Lustre WAN OSS and OST hardware and specifications.

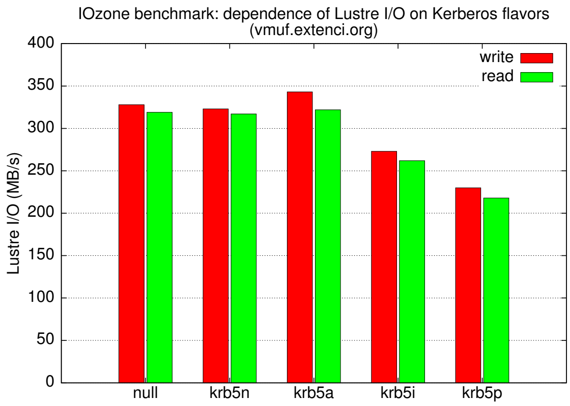


Figure 4. Impact of Kerberos flavors on I/O rates on the LAN.

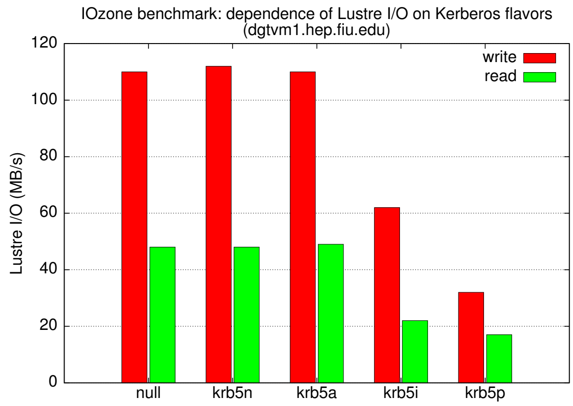


Figure 5. Impact of Kerberos flavors on I/O rates on the WAN.

WAN, where the latencies can be tens of milliseconds. We use krb5n, which has very minor overhead, for all subsequent tests.

OST pools enable the grouping of OSTs for file striping purposes with automated free-space leveling occurring within the pool. OST pool usage is specified and stored along with other striping information such as stripe count and size for directories, or individual files. IOzone benchmarks on UF and Fermilab OST pools show marked improvement on the Lustre I/O when local storage is used. Tier-3 universities (FIU, USF, FSU) and Fermilab can use the closest OST

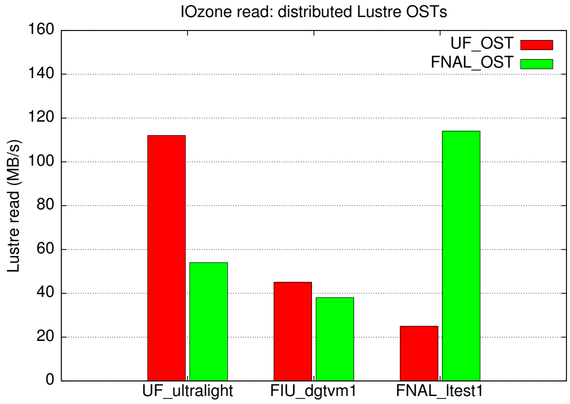


Figure 6. Comparison of I/O rates for reading from local or remote OST pools.

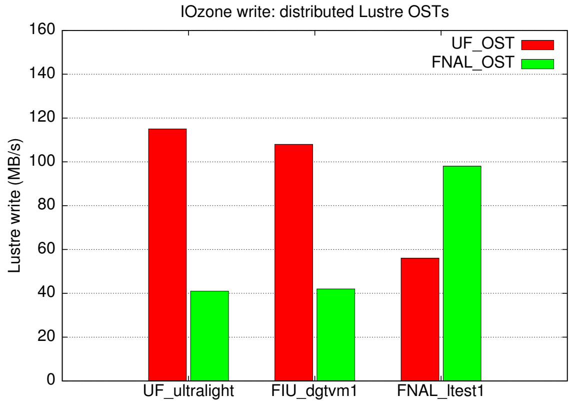


Figure 7. Comparison of I/O rates for writing to local or remote OST pools.

pool available. OST pools created at UF and Fermilab contribute to a total storage of 60 TB.

The I/O rates for the distributed OST pools at the University of Florida and FNAL (as determined by IOzone) depend on the distance from the client to the server: best for LAN, worst for WAN over longer distances. The round trip times (RTT) are: UF-FNAL 47 ms, UF-FIU 14 ms, FIU-FNAL 51 ms. The results are summarized in Figures 6 and 7.

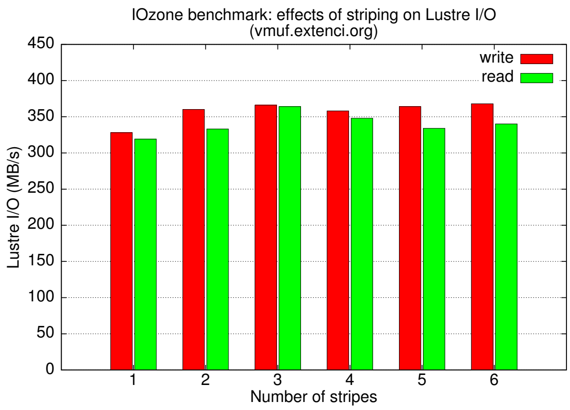


Figure 8. Impact of striping on I/O rates over the LAN.

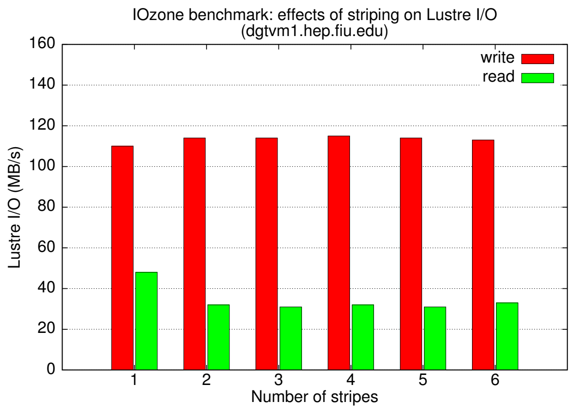


Figure 9. Impact of striping on I/O rates over the WAN.

We have studied the effects of striping of files on the performance and have observed no tangible benefits, as shown in Figures 8 and 9. As striping increases the risk of data loss, in all subsequent tests we have used a stripe size of one (no striping).

To test the scalability of our system we have performed IOzone runs with a realistic mix of clients accessing data over the LAN and the WAN. First we ran sets of tests sequentially (stacked) in order to determine the best performance of our system for different client/server combinations. Then we ran all clients in parallel. As shown on Figure 10, the simultaneous run results in I/O rates close to the stacked rates.

From the CMS detector, the raw data goes online to storage at CERN, a Tier0 site. Here the raw, full event information is repacked and reconstructed (RECO), and unsorted streams are sorted into physics streams of events with similar characteristics. Selection algorithms produce the AOD (Analysis Object Data) and RAW, RECO and AOD are exported to Tier1 sites. At

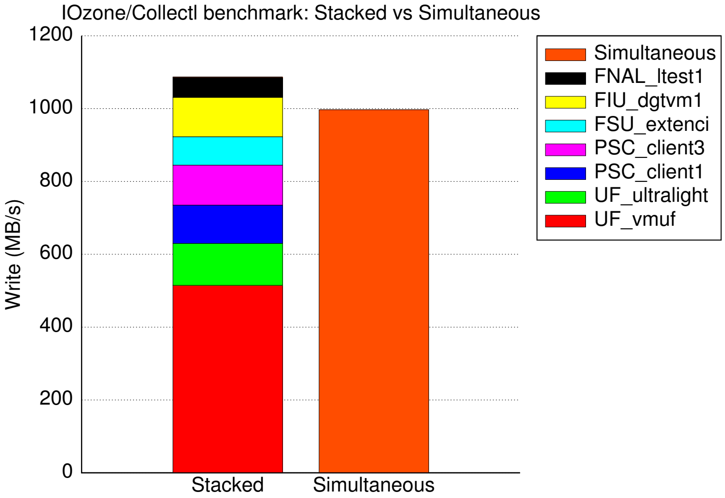


Figure 10. Test of the scalability of the distributed Lustre filesystem.

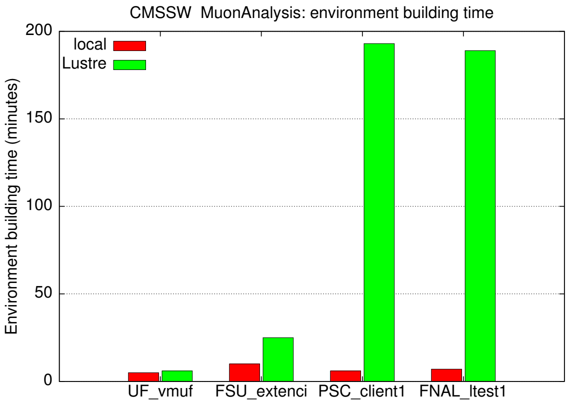


Figure 11. CMSSW SCRAM building time on Lustre and Local partitions. Compilation and link (involving many small files) time is good over the LAN or for close clients: FSU (RTT 4 ms), much slower for far clients: FNAL and PSC (RTT 46-47 ms).

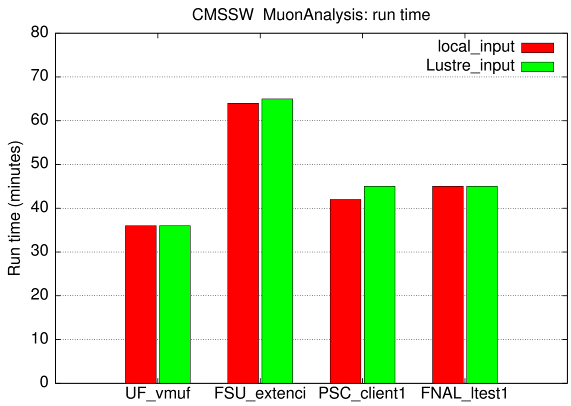
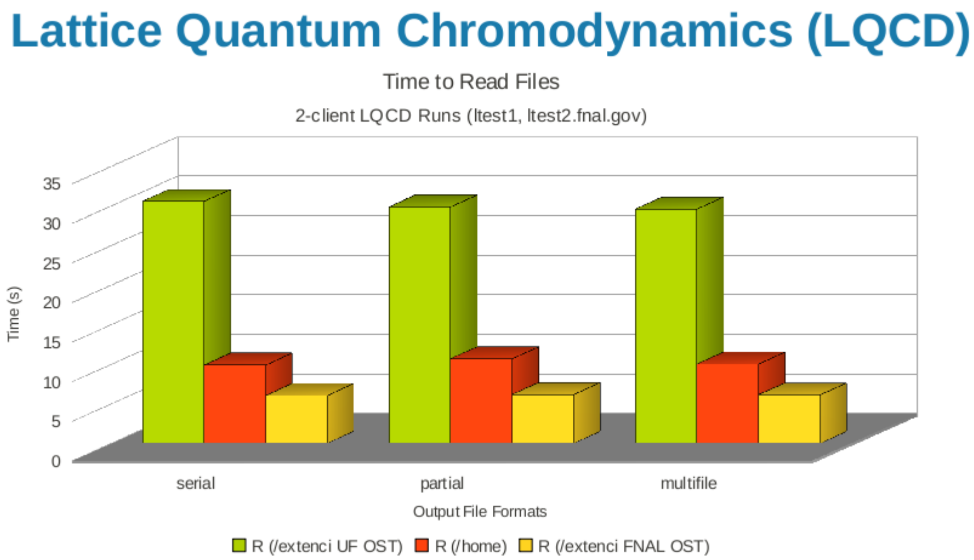


Figure 12. CMSSW run time on real muon data using Lustre and Local partitions.

the Tier2 level, Monte Carlo events are generated, detector interactions are simulated, events are reconstructed, and moved into disk storage for later use. At the Tier2/Tier3 stage, users prepare the analysis code, run it on the experimental data and finally collect results.

There are two stages in the CMS data analysis - SCRAM and RUN. SCRAM involves the CMS environment building time for the compilation and linking of various libraries; RUN consists of actual data processing. We tested CMSSW_3_9_7 muon analysis code at four remote Lustre clients. Preliminary results of CMS data processing on the Lustre WAN gave poor client performance at remote sites like PSC and FNAL. Network latency greatly lengthened the roundtrip time for every access back to the servers. CMS SCRAM encountered over 8.5×10^4 file accesses (2.4×10^4 opens, 2.4×10^4 stats, 3.6×10^4 lstats and 1.5×10^3 readlinks) from 2×10^4 directories holding 1.8×10^5 files with sizes ranging from a few KB to several MB. The LOSF (lots of small files) scenario compounded the effects of large network latencies from PSC or FNAL during CMSSW SCRAM, see Figure 11. As a result, SCRAM's environment building time for CMS was adversely affected. What locally took a minute or two for SCRAM to complete, lengthened to 200 min (3 hrs) at more remote sites. Making the 12 GB CMSSW software resident on the Lustre VM client's local partition, `/local/cmssw`, while placing the input or data files on the remote Lustre filesystem solved this problem.

The picture changes considerably in the run stage, especially for more CPU intensive tasks. In our CMSSW tests, runs performing reconstruction over a 2 GB data file gave comparable run times for accessing locally or remotely stored data files, as shown in Figure 12.



Number of Input Files	Number of Output Files	Input File Sizes	Output File Sizes	Output File Formats
2	1	579MB, 495B	579MB	serial
2	16	"	37MB	partial
2	16	"	37MB	multifile

Figure 13. Lattice QCD benchmarks: read times; shorter is better. A distributed Lustre filesystem with OSS at UF and OST pool at FNAL is used.

The QCD theory describes the interactions between quarks and gluons. In lattice QCD, the fields representing the quarks are defined at lattice sites while the gluon fields are defined on the links connecting the neighboring sites. This approximation approaches continuum QCD as the spacing between lattice sites is reduced to zero. The software stack used to build the static SU3 for numerical simulations of 4D lattice gauge theory includes milc_qcd-7.6.3, mvapich-2.1.7-r5225, scidac and scidac-mvapich.

LQCD was ran from two clients at Fermilab having the Infiniband (IB) network back-end for message-passing while performing I/O on Lustre. Three filesystems are compared - UF OST pool, Fermilab's client local partition (`/tmp, /home`), and Fermilab's OST pool. Results indicate the benefit of using Fermilab's local OSTs (versus UF's OSTs) with the 6-fold improvement on read I/O across all output file formats in contrast to the 3-fold I/O improvement using the local partition. Results for write I/O, however, showed preference for the local partition. The results are summarized in Figures 13 and 14.

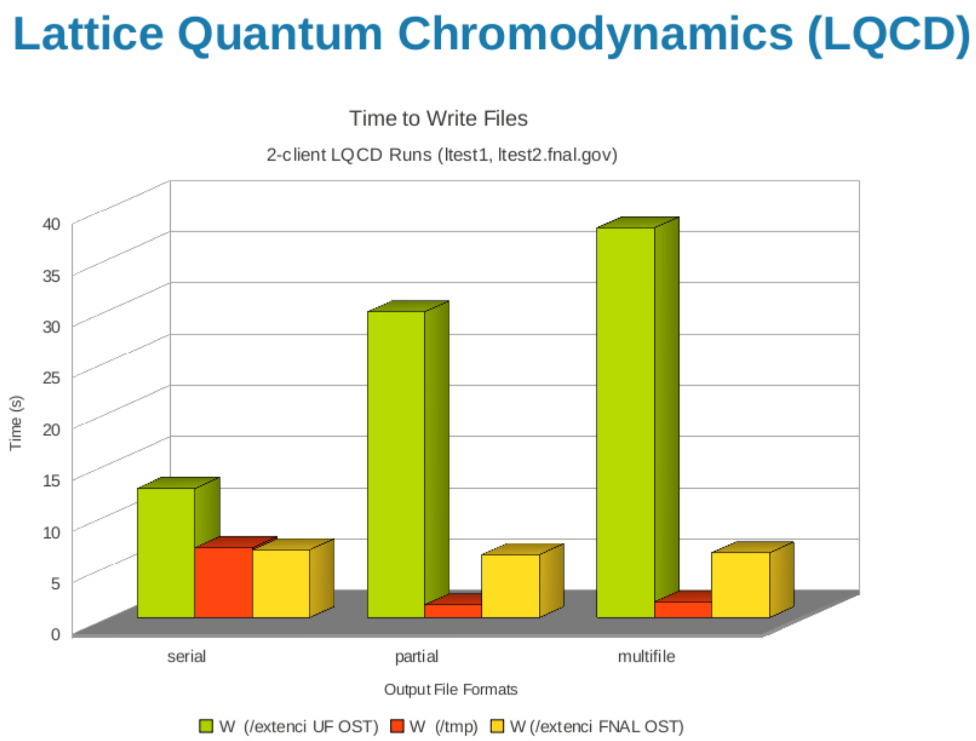


Figure 14. Lattice QCD benchmarks: write times; shorter is better. A distributed Lustre filesystem with OSS at UF and OST pool at FNAL is used.

4. Issues

The dominating issue we encountered was the lack of support for Kerberos in Lustre. The clients sometimes crash, leading to instabilities when running the applications, in effect slowing down our progress. Fortunately, work is ongoing now to include Kerberos functionality into WhamCloud's test suite and finally into the Lustre mainstream. This will help to move our test bed in stable production status.

5. Outlook

We have successfully built and deployed a Kerberos based secure Lustre filesystem, distributed over the WAN, with 60 TB of storage located in pools at the University of Florida and Fermilab. It is accessed through KDC, MDS and OSS servers located at the University of Florida. Similar setups can be used for easy, convenient and secure user access to large amounts of data, for example in an ecosystem like the Florida CMS Tier2 center and several Florida Tier3 centers associated with it. Our test results demonstrate good I/O performance and will be expanded with more use cases.

We continue our efforts in profiling and tuning of applications in order to optimize their performance on the WAN, to fully understand the interplay between CMS data tree structures and performance, and to predict which types of files would give best I/O rates on the distributed Lustre filesystem.

We hope to better integrate with CERN tools and the OSG and XSEDE resources to increase the project's interoperability with other organizations. Having a functional PKINIT using grid certificates would also allow users to authenticate in the Kerberos realm using their OSG X509 certificates.

Acknowledgments

We acknowledge the funding from the National Science Foundation, NSF Proposal 1007115 ExTENCI with the OSG.

References

[1] The Lustre filesystem. Available at `http://www.lustre.org` .

[2] The Lustre 2 filesystem. Available at `http://wiki.whamcloud.com/display/PUB/documentation` .

[3] MIT Kerberos. Available at `http://web.mit.edu/kerberos` .

[4] The CMS experiment. Available at `http://cms.web.cern.ch/content/cms-physics` .

[5] Lattice QCD. Available at `http://www.physics.utah.edu/~detar/milc/milc_qcd.html` .

[6] Palencia J *et al.* "Using Kerberized Lustre over the WAN for High Energy Physics Data," Proceedings of the Lustre Users Group 2012 Meeting, Austin, TX, April 2012. Available at `http://www.opensfs.org/wp-content/uploads/2011/11/lug2012-v20.pdf` .